*Nucleic Acids Research, 2007, Vol. 35, Web Server issue* ***W585–W587***
*doi:10.1093/nar/gkm259*

# WoLF PSORT: protein localization predictor

**Paul Horton[1], Keun-Joon Park[1,2], Takeshi Obayashi[3], Naoya Fujita[1,3], Hajime Harada[1], C.J. Adams-Collier[4] and Kenta Nakai[3,*]**

[1]Computational Biology Research Center, AIST, Tokyo, Japan, [2]Center for Genome Science, National Institute of Health, Korea Center for Disease Control & Prevention, 5 Nokbeon-Dong, Eunpyung-Gu, Seoul 122-701 Korea, [3]Human Genome Center, Institute of Medical Science, University of Tokyo, Tokyo, Japan and [4]Collier Technologies, Everett, WA, USA

Received January 30, 2007; Revised March 26, 2007; Accepted April 8, 2007

## ABSTRACT

**WoLF PSORT is an extension of the PSORT II program for protein subcellular location prediction. WoLF PSORT converts protein amino acid sequences into numerical localization features; based on sorting signals, amino acid composition and functional motifs such as DNA-binding motifs. After conversion, a simple *k*-nearest neighbor classifier is used for prediction. Using html, the evidence for each prediction is shown in two ways: (i) a list of proteins of known localization with the most similar localization features to the query, and (ii) tables with detailed information about individual localization features. For convenience, sequence alignments of the query to similar proteins and links to UniProt and Gene Ontology are provided. Taken together, this information allows a user to understand the evidence (or lack thereof) behind the predictions made for particular proteins. WoLF PSORT is available at wolfpsort.org**

## INTRODUCTION

Bilipid membranes divide eukaryotic cells into various types of organelles containing characteristic proteins and performing specialized functions. Thus, subcellular localization information gives an important clue to a protein's function. Although localization signals in mRNA appear to play some role (1), the main determinant of a protein's localization residues in the protein's amino acid sequence. (We recommend wikipedia.org/wiki/Protein_targeting for a brief overview and Alberts *et al.* (2) for a textbook description.)

Numerous experiments to determine protein localization have been performed to date. These can broadly be classified as: small-scale experiments—the results of which continue to accumulate in public databases, such as UniProt (3) and Gene Ontology (4); and large-scale experiments using epitope (5) or green fluorescent protein (GFP) (6) tagging, or by separation of organelles by centrifugation combined with protein identification by mass spectrometry (7,8).

Although they provide invaluable information, the coverage of experimental data is only high for model organisms, particularly yeast. Moreover, the agreement amongst large-scale experimental data is only 75–80% (6–9). Thus, computational prediction of localization from amino acid remains an important topic.

Numerous computational methods are available [reviewed in (10,11)]. Some (including WoLF PSORT) have recently been benchmarked by Sprenger *et al.* (12), who found the computational methods to be useful for sites, such as the nucleus, for which many training examples can be easily obtained from UniProt (which is the source of most or all of the training data for most prediction methods—including WoLF PSORT). The different methods they benchmarked were found to have different strengths. Here, we describe the public server for our WoLF PSORT method.

## PREDICTION METHOD

WoLF PSORT is an extension of PSORT II (13,14) and also uses the PSORT (15) localization features for prediction. In addition, WoLF PSORT uses some features from iPSORT (16) and amino acid composition. Those features are used to convert amino acid sequences into numerical vectors, which are then classified with a weighted *k*-nearest neighbor classifier. WoLF PSORT uses a wrapper method to select and use only the most relevant features. This reduces the amount of information which needs to be considered (and displayed) for the user to interpret individual predictions and may also make the predictor less prone to over learning. The prediction method has described in more detail elsewhere (17).

*To whom correspondence should be addressed. Tel: +81-3-5449-5131; Fax: +81-3-5449-5133; Email: knakai@ims.u-tokyo.ac.jp

32 Nearest Neighbors

| id | site | distance | identity | comments |
|---|---|---|---|---|
| TCOF_HUMAN | nucl | 0.257022 | 94.8253% | [Uniprot] SWISS-PROT45:Nuclear, nucleolus. GO:0005730; C:nucleolus; Evidence:IDA. |
| ABL1_HUMAN | cyto_nucl | 1341.71 | 14.8522% | [Uniprot] SWISS-PROT45:Cytoplasmic. WoLF PSORT custom: Shuttling cytoplasm and nu |
| ENL_HUMAN | nucl | 2316.82 | 11.7608% | [Uniprot] SWISS-PROT45:Nuclear. GO:0005634; C:nucleus; Evidence:TAS. |
| YEMA_DROME | nucl | 2651.38 | 15.457% | [Uniprot] SWISS-PROT45:Nuclear. |
| GOGA4_MOUSE | cyto | 2842.74 | 14.1197% | [Uniprot] SWISS-PROT45:Cytoplasmic; peripheral membrane protein associated with t |
| CPD1_DROME | nucl | 3153.91 | 8.26613% | [Uniprot] SWISS-PROT45:Nuclear. |
| FYB_MOUSE | cyto_nucl | 3191.38 | 14.5161% | [Uniprot] SWISS-PROT45:Nuclear and cytoplasmic. |

**Figure 1.** Part of the list of proteins similar to the query protein, an isoform of TCOF_HUMAN, is shown. For each neighbor the following is shown: UniProt ID, localization site, the distance in localization features from the query, the percent identity to the query, a link to its UniProt entry, the subcellular localization line from UniProt and other available localization information.

### Dataset

The WoLF PSORT dataset is divided into fungi, plant and animal containing 2113, 2333 and 12771 proteins, respectively. The current data was primarily obtained from UniProt (3) version 45, but subcellular localization information from Gene Ontology (4) was also used. Entries with evidence codes {TAS, IDA, IMP} were included, with manual revisions in a few cases. We intend to update these datasets regularly in the future.

## LOCALIZATION SITES AND PREDICTION ACCURACY

WoLF PSORT classifies proteins into more than 10 localization sites, including dual localization such as proteins which shuttle between the cytosol and nucleus. Based on our cross-validation studies (17), we estimate sensitivity and specificity of around 70% for: *nucleus*, *mitochondria*, *cytosol*, *plasma membrane*, *extracellular* and (in plants) *chloroplast*. For other sites, such as peroxisome, Golgi, etc. the sensitivity is very low, but useful predictions are still made in some cases. For example, the *Arabidopsis* seed protein 12S1_ARATH is reasonably predicted to localize to the vacuole even though only one of its neighbors (see below) shares significant sequence similarity. An independent test (12) on mouse proteins gave a significantly lower estimate of WoLF PSORT's prediction accuracy (around 50%). This discrepancy may be explained by the over-representation of well-studied proteins in the WoLF PSORT training data and perhaps also by the size of their test data (in particular, their 'LOC2145' test set contained only 87 cytosolic proteins) or differences in site definition.

## PREDICTION RESULTS DISPLAY

The *k*-nearest neighbors classifier allows for an intuitive display of the prediction results which is exactly analogous to sequence similarity search. Using multifasta format, multiple sequences can be given in a query. The first page returned from the server gives a one line summary of the result for each query sequence. For example the prediction summary line for the TCOF_HUMAN protein is:

```
TCOF_HUMAN details nucl: 27.5, cyto_nucl: 17, cyto:
3.5, extr: 1
```

The localization sites are abbreviated to four letter codes (documented on the server) with dual localization denoted by joining the four letter codes with an underscore character. The numbers roughly indicate the number of nearest neighbors to the query which localize to each site—but are adjusted to account for the possibility of dual localization (17).

### Neighbor list

Details about the queries neighbor list and localization signals can be obtained by following the 'details' link. The first part of the display page is a neighbor list table such as the one shown in Figure 1. This list gives information regarding the query's neighbors (proteins in the WoLF PSORT training data that have the most similar localization features). For user convenience, the percent identity and a link to the alignment of each neighbor to the query is given. Sequence similarity is not used for prediction but can provide additional corroborating evidence in many cases. Links to the relevant entries in UniProt, Gene Ontology and TAIR (www.arabidopsis.org) for many *Arabidopsis* entries are also provided.

### Localization feature table

By scrolling down on the detailed results pages, one can find a feature table giving the values of each localization feature for the query and its neighbors. In some cases, the individual values can help support (or question) the predicted site. For example in the case of TCOF_HUMAN (Figure 2), the 99 percentile value of the PSORT localization feature 'nuc' (which is based on nuclear localization signals and DNA-binding site motifs), is consistent with the nuclear prediction. Below the normalized table, a similar table with the raw feature values is displayed.

## IMPLEMENTATION

The server is implemented with Mason (www.masonhq.com), which allows convenient embedding of logic and computed results into html via the Perl programming language. Multiple requests are handled with the simple strategy of returning the results in a URI containing an MD5 hash of the query contents.

**Normalized Feature Values**

| id | site | iPSORT | | PSORT Features | | | | | | | | | | | | | | Amino Acid Content | | | | Misc. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | -1 25 | MxHy1 30 | act | alm | dna | gvh | leu | mNt | mip | mit | myr | nuc | rib | rnp | tms | tyr | vac | C | I | K | S | length |
| TCOF_HUMAN | nucl? | 87 | 51 | 50 | 73 | 44 | 48 | 46 | 49 | 26 | 67 | 49 | 99 | 50 | 50 | 27 | 49 | 48 | 7 | 3 | 94 | 98 | 98 |
| TCOF_HUMAN | nucl | 87 | 51 | 50 | 73 | 44 | 48 | 46 | 49 | 26 | 67 | 49 | 99 | 50 | 50 | 27 | 49 | 48 | 7 | 4 | 94 | 98 | 98 |
| ABL1_HUMAN | cyto_nucl | 87 | 60 | 50 | 65 | 44 | 37 | 46 | 49 | 26 | 54 | 49 | 97 | 50 | 50 | 27 | 49 | 48 | 24 | 20 | 75 | 89 | 96 |
| ENL_HUMAN | nucl | 87 | 15 | 50 | 89 | 44 | 36 | 46 | 49 | 26 | 42 | 49 | 95 | 50 | 50 | 27 | 49 | 48 | 24 | 5 | 93 | 99 | 80 |
| YEMA_DROME | nucl | 78 | 42 | 50 | 58 | 44 | 41 | 46 | 49 | 65 | 45 | 49 | 97 | 50 | 50 | 27 | 49 | 48 | 11 | 29 | 89 | 96 | 94 |
| GOGA4_MOUSE | cyto | 78 | 22 | 50 | 92 | 44 | 45 | 46 | 49 | 26 | 49 | 49 | 85 | 50 | 50 | 27 | 49 | 48 | 21 | 13 | 90 | 66 | 99 |
| CPD1_DROME | nucl | 93 | 43 | 50 | 85 | 44 | 23 | 46 | 49 | 26 | 25 | 49 | 97 | 50 | 50 | 27 | 49 | 48 | 3 | 7 | 92 | 94 | 51 |
| FYB_MOUSE | cyto_nucl | 64 | 17 | 50 | 92 | 44 | 34 | 46 | 49 | 26 | 16 | 49 | 96 | 50 | 50 | 27 | 49 | 48 | 9 | 19 | 95 | 74 | 91 |
| SFRC_RAT | nucl | 46 | 64 | 50 | 57 | 44 | 66 | 46 | 49 | 26 | 63 | 49 | 99 | 50 | 50 | 27 | 49 | 48 | 8 | 14 | 96 | 98 | 73 |
| SFRC_HUMAN | nucl | 46 | 59 | 50 | 58 | 44 | 76 | 46 | 49 | 26 | 50 | 49 | 99 | 50 | 50 | 27 | 49 | 48 | 6 | 21 | 97 | 97 | 75 |

**Figure 2.** The localization features for the query and its neighbors are shown. The values are normalized to percentiles relative to the WoLF PSORT training data. Neighbor values shown in blue are within 10% points to the query value, while those shown in red are 20 or more percentile point different from the query.

Upon sending a query a wait page is shown, followed by an automatic redirect to the results page upon task completion (usually requiring around 40 s). Task scheduling is delegated to Apache and the Linux operating system. Multiple sequences are allowed in one query, but we currently limit the query size to 64 KB. For large-scale use, such as whole genome annotation, we encourage users to download the stand-alone package (available on the server) and run WoLF PSORT locally.

## SUMMARY

WoLF PSORT not only provides subcellular localization prediction with competitive accuracy, but also provides detailed information relevant to protein localization to help users to form their own hypotheses.

## ACKNOWLEDGEMENTS

KN was partly supported by a grant from the National Project on Protein Structural and Functional Analyses by the Ministry of Education, Culture, Sports, Science and Technology in Japan. The annual budget of the Human Genome Center was used for the publication of this paper.

*Conflict of interest statement*. None declared.

## REFERENCES

1. Gonsalvez,G.B., Urbinati,C.R. and Long,R.M. (2005) RNA localization in yeast: moving towards a mechanism. *Biol. Cell*, **97**, 75–86.
2. Alberts,B., Bray,D., Lewis,J., Raff,M., Roberts,K. and Watson,J.D. (2002) *Molecular Biology of the Cell, 4th edn*. Garland Publishing. New York.
3. Bairoch,A., Apweiler,R., Wu,H., Barker,C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R. *et al.* (2005) The universal protein resource (UniProt). *NAR*, **33**, D154–D159.
4. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., *et al.* (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.
5. Kumar,A., Agarwal,S., Heyman,J.A., Matson,S., Heidtman,M., Piccirillo,S., Umansky,L., Drawid,A., Jansen,R., *et al.* (2002) Subcellular localization of the yeast proteome. *Genes Dev.*, **16**, 707–719.
6. Huh,W.-K., Falvo,J.V., Gerke,L.G., Carroll,A.S., Howson,R.W., Weissman,J.S. and O' Shea,E.K. (2003) Global analysis of protein localization in budding yeast. *Nature*, **425**, 686–691.
7. Prokisch,H., Scharfe,C., Camp II,D.G., Xiao,W., David,L., Andreoli,C., Monroe,M.E., Moore,R.J., Gritsenko,M.A., *et al.* (2004) Integrative analysis of the mitochondrial proteome in yeast. *PLoS Biol.*, 2(**6**): e160.
8. Foster,L.J., de Hoog,C.L., Zhang,Y., Xie,X., Mootha,V.K. and Mann, M. (2006) A mammalian organelle map by protein correlation profiling. *Cell*, **125**, 187–199.
9. Nair,R. and Rost,B. (2005) Mimicking cellular sorting improves prediction of subcellular localization. *JMB*, **348**, 85–100.
10. Emanuelsson,O. (2002) Predicting protein subcellular localisation from amino acid sequence information. *Brief. Bioinformatics*, **3**, 361–376.
11. Horton,P., Mukai,Y. and Nakai,K. (2004) Protein localization prediction. In Wong,L. (ed.), *The Practical Bioinformatician*, Chapter 9, pp. 193–215, World Scientific 5 Toh Tuck Link, Singapore 596224.
12. Sprenger,J., Fink,J.L. and Teasdale,R.D. (2006) Evaluation and comparison of mammalian subcellular localization prediction methods. *BMC Bioinformatics*, **7**(Suppl 5), S3.
13. Horton,P. and Nakai,K. (1997) Better prediction of protein cellular localization sites with the k nearest neighbors classifier. In Gaasterland,T., Karp,P., Karplus,K., Ouzounis,C., Sander,C. and Valencia,A. (eds), *Proceeding of the Fifth International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, Halkidiki, Greece, pp. 147–152.
14. Nakai,K. and Horton,P. (1999) Psort: a program for detecting sorting signals in proteins and determining their subcellular localization. *TIBS*, **24**, 34.
15. Nakai,K. and Kanehisa,M. (1992) A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics*, **14**, 897–911.
16. Bannai,H., Tamada,Y., Maruyama,O., Nakai,K. and Miyano,S. (2002) Extensive feature detection of N-terminal protein sorting signals. *Bioinformatics*, **18**, 298–305.
17. Horton,P., Park, K.-J., Obayashi,T. and Nakai,K. (2006) Protein subcellular localization prediction withWoLF PSORT. In Jiang,T., Yang,U.-C. and Chen,Y.-P.P. (eds), *Proceedings of the 4th Annual Asia Pacific Bioinformatics Conference, APBC06*, Imperial College Press, London, pp. 39–48.